Atypical Queries in eCommerce

Neeraj Pradhan
Search and Data Mining
Groupon
npradhan@groupon.com

Vinay Deolalikar
Search and Data Mining
Groupon
vdeolalikar@groupon.com

Kang Li
Search and Data Mining
Groupon
kli@groupon.com

ABSTRACT

Understanding how specific, ambiguous, or broad the intent of a search query is, across all users of the system, is important in improving search relevance in eCommerce. There is scant literature on such a structural characterization of queries in eCommerce. In this paper, we use query-click log data to address the problem of identifying "atypical queries": these are queries that are extremal in terms of specificity, ambiguity, or breadth of intent. We isolate three components of atypicality: geometric, statistical, and topological. We demonstrate, using query-click logs at Groupon, that certain *combinations* of these properties render a query atypical, and discuss how search analysts treat such queries differently. Our work is being used to improve search relevance at Groupon.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process; relevance feedback*

Keywords

eCommerce search; atypical queries; query intent disambiguation

1. INTRODUCTION

eCommerce is one of the cornerstones of the internet age. As of 2013, the global eCommerce business-to-consumer market was valued at USD 1.3 Trillion. eCommerce companies devote a large amount of effort in trying to gain structural insights about user intent for high-volume queries. In particular, they are highly interested in the specificity, breadth, and ambiguity associated with a query. Understanding these properties of a query allows them to show an appealing result set to their users. It also enables them to personalize this result set based on user attributes. Finally, such an understanding powers alternative product recommendations made to users who have issued the query.

Although characterizing queries in the manner described above requires intensive effort, often with a significant manual component, there is scant literature on this problem. In contrast, the problem of understanding query ambiguity (and its relationship to query performance) is an active field in text document retrieval.

In this paper, we address this gap. We consider the research question of automatically identifying queries that are extremal in each of the aforementioned aspects: namely, queries that are highly specific, or very broad, or highly ambiguous. We call such queries *atypical queries*. Furthermore, we would like to identify atypical queries with very little manual intervention or inspection, except for an initial configuration of our system.

The hurdle to answering our research questions lies in extracting notions of typicality that are *general* enough to capture aggregate user behavior. Our approach is to *bootstrap out notions of typicality from query-click data, without external knowledge*. This makes our methods completely general. We demonstrate our approach on a dataset of six months of query logs from the Groupon site.

Our contributions are outlined below.

1. We introduce research on a broad structural understanding of queries in eCommerce.
2. We provide a theoretical framework for identifying queries that are atypical in specificity, breadth, or ambiguity.
3. Our framework considers three aspects of a query-click distribution: geometric, statistical, and topological. We show that certain *combinations* of these aspects render a query atypical. We provide examples of atypical queries.
4. The three properties above, being inherent to the query-click distribution, represent structural knowledge that can be used more generally.

2. KEY IDEA

Users interact with eCommerce sites primarily through searches. A specific search is called a *query* and may consist of one or more terms (e.g. "adidas" or "Apple iMac"). Each product in the inventory is placed into one or more *categories* in the product taxonomy (e.g. men's shoes or electronics, for the previous examples). Let us assume that we are studying the set of all queries $\mathscr{Q} = \{Q_1, Q_2, \dots, Q_m\}$ made over a specific period of time. Let the set of categories in the product taxonomy be $\mathscr{C} = \{C_1, C_2, \dots, C_n\}$.

When a user makes a query, and then clicks on products in the result set, these clicks are recorded against the categories that the products are in. For each query, we aggregate these user clicks over a certain time period. These aggregated clicks, when normalized, form a distribution over categories. For example, the query "vintage" might result in a bimodal distribution, with peaks over the categories "wine stores" and "auto parts." Some formal definitions are in order.

Definition 1. The *signature* of a query $Q_i \in \mathscr{Q}$, is denoted by $\text{sign}(Q_i)$, and defined as the vector $\mathbf{w}_i = (w_{i1}, \dots, w_{in})$, where w_{ij} is the ag-

CIKM'15, October 19–23, 2015, Melbourne, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806605 .

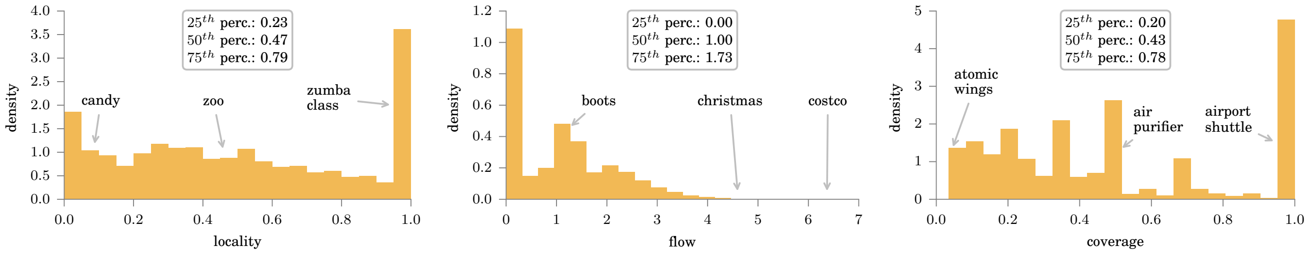

Figure 1: Distributions of locality, flow, and coverage for queries in Groupon_QL. Also shown are quartile values, and examples of some queries and where they lie in the distribution.

gregate number of clicks on (products of) category C_j, generated by query Q_i.

The signature $\text{sign}(Q_i)$ can be converted into a probability distribution over $\mathscr{C}$ by normalizing by $\sum_{j=1}^{j=n} w_{ij}$. We refer to this as the *normalized signature*. To keep notation light, we use the same symbols for the signature and the normalized signature.

Definition 2. The *support of a query Q_i* is defined as

$$\text{supp}(Q_i) = \{C_j \colon w_{ij} \neq 0\}.$$

Namely, it is the set of categories where the query Q_i has a non-zero signature.

Definition 3. The set of signatures for all $Q \in \mathscr{Q}$ is called the *query-category click model*.

The key idea that guides our approach is the following.

> Certain combinations of properties of the query's signature, relative to that of the query-category click model as a whole, mark the query as being atypical.

The question now is: what are these properties of the query's signature? Based on our extensive collaboration with search analysts at Groupon, we have identified three properties that we call *locality, flow, and coverage*. These are formalized in the next section.

3. THEORETICAL FRAMEWORK

In this section, we build a mathematical framework in which the ideas of the previous section can be precisely expressed.

3.1 Signatures, Similarity, and Closure

In order to define the quantities of our interest, it will help to place some preliminary definitions.

Definition 4. Let w_{ij} be the aggregate number of clicks generated by Q_i that go to category C_j in the query-category click model (Def. 3). The *signature of a category C_j* is defined as

$$\text{sign}\,(C_j) = (w_{1j}, \ldots, w_{mj}).$$

Definition 5. The *similarity* between two categories C_i and C_j is defined as the cosine similarity of their query signatures.

$$\text{sim}\,(C_i, C_j) = \cos\left(\text{sign}(C_i), \text{sign}(C_j)\right).$$

Definition 6. The *closure* of a category C_i, w.r.t. a threshold T is defined as follows:

$$\text{closure}\,(C_i) = \{C_j \colon \text{sim}(C_i, C_j) > T\}.$$

The value of T is determined empirically based on the application, and is suppressed in order to keep notation light.

3.2 Locality, Flow, and Coverage

Locality (Geometric).

Definition 7. The *locality of a query* is the average of the pairwise similarities of the categories in its support. Namely,

$$\text{locality}(Q) = \text{avg}\{\text{sim}(C_i, C_j) \colon C_i, C_j \in \text{supp}(Q) \text{ and } C_i \neq C_j\}.$$

If $|\,\text{supp}\,(Q)| = 1$, then $\text{locality}(Q) = 1$ by definition.

Locality is a geometric notion: it measures *distances* between the categories in $\text{supp}\,(Q)$. These distances are, in turn, based on the aggregate query-click log. Namely, two categories are "close" when many queries have clicks that distribute over both categories.

Flow (Statistical).

We refer the reader to [2] for a detailed treatment of entropy.

Definition 8. The (information-theoretic) *entropy* of a discrete probability distribution P, whose masses are $\{p_i\}$, is defined as

$$\text{entropy}(P) = -\sum_i p_i \log_2(p_i).$$

Definition 9. The *flow* for a query is defined as the entropy of its normalized signature.

$$\text{flow}(Q) = \text{entropy}(\text{sign}(Q)).$$

Flow is a statistical notion: it measures how "evenly" the distribution over categories in $\text{supp}\,(Q)$ is. Namely, what is the uncertainty (in the sense of information theory) in this distribution.

Coverage (Topological).

Definition 10. The *coverage* of a query Q is defined as

$$\text{coverage}(Q) = \frac{|\,\text{supp}(Q)|}{|\,\text{closure}(\text{supp}(Q))|}.$$

Coverage is related to the notion of local cover in topology [7]. It examines how closed $\text{supp}\,(Q)$ is w.r.t. "similar" categories. Namely, if $C_i \in \text{supp}\,(Q)$, then how many categories similar to C_i are also in $\text{supp}\,(Q)$.

4. REGIONS OF ATYPICALITY

In this section, we argue that there are certain combinations of locality, flow, and coverage that mark a query as being atypical.

Our framework is general, and the set of atypical combinations can be expanded for specific applications. We provide three such combinations as working examples, explain why they are atypical, and what eCommerce search analysts aim to do in each case. §6.4 gives examples of queries in each of the three regions.

4.1 Region I: Low Locality and High Flow

Locality measures how similar are the categories that appear in the signature of a query. Low locality means that the clicks generated by a query are distributing over dissimilar categories. High flow means that there is a more uniform distribution over several such categories. Due to this, it is more likely that the query is broad, having inventory in multiple categories. For such queries, search analysts aim to **maximize recall** so that all products that could potentially match the broad query be shown.

4.2 Region II: Low Locality and Low Flow

This combination of properties indicates that the query is being made by a small number of different sets of users having different intents. A toy example would be a query such as "apple" which might be made with intents varying between food and computers. Food and computers have a low similarity score since they do not frequently co-occur in query signatures; consequently, such a query would have low locality. Such queries are **candidates for disambiguation** by eCommerce search analysts.[1]

4.3 Region III: High Locality, Low Flow, and Low Coverage

The combination of properties in this case indicate high specificity of the query. Not only does the signature not have a large support, the support is very "selective" in that it has low coverage, and high locality. eCommerce search analysts treat such queries as being highly specific, and aim for **high precision** here in order to satisfy the user. Ensuring high recall is secondary for such queries.

5. RELATED WORK

To our knowledge, this is the first work focused on the question of characterizing atypical queries in the context of eCommerce. There are a few related areas of research in information retrieval that our work touches upon.

The problem of predicting query performance has been studied by many research groups in the context of content-based queries [3, 1, 10] or, more generally, web search queries [11, 6]. The research can broadly be categorized as being focused on pre-retrieval or post-retrieval performance measures. Pre-retrieval measures predict how a query would perform in a search system by using only the properties associated with the query and assuming that the result-set is not available. A survey of pre-retrieval performance predictors can be found in [5]. We focus specifically on post-retrieval measures as the low dimensionality of the query space in eCommerce negates many of the advantages of pre-retrieval performance predictors. Some examples of post-retrieval measures are Jensen-Shannon divergence [1] and weighted information gain [11], which predict the performance of a query in terms of average precision on a retrieval task, using characteristics of the query and the result-set.

A related area of research is the problem of detecting ambiguous queries. In [3], the authors claim that a strong predictor of low performing queries is lack of query ambiguity, namely how much information does the query term contain with regard to specifying user intent. They have quantified this through a measure called clarity score which measures the relative entropy between the query and the retrieved result-set. The variability in user intent, captured by query ambiguity measures, has also been studied in the context of search personalization by [9, 4]. Dou et al. [4] introduce the click entropy measure for a query to judge its suitability for personalization. This is one of the measures that we use in the classification of atypical queries. Song et al. [8] developed a supervised learning model to classify ambiguous queries and recognized that queries that are not ambiguous could either be "broad" or "clear" queries. Their model, however, does not distinguish between these two classes. Our notion of atypicality in the context of eCommerce is broader. Rather than defining a single measure of ambiguity, we focus on measures derived from the implicit query-click data that can be collectively used to categorize queries in terms of the broadness or ambiguity of user intent. Our framework is general and can be adapted to any domain-specific search engine.

6. EMPIRICAL WORK

6.1 Dataset

We demonstrate our approach on a dataset, Groupon_QL, containing roughly 41,000 distinct queries with the highest query volume, aggregated from search logs at Groupon over a six month period beginning October 2014. For each query, the aggregated user clicks are normalized to form a distribution over the set of categories. The number of categories is 986. In order to reduce noise in the computation of locality, we only consider categories having at least 10% contribution in the normalized query signature.

6.2 Distribution of Key Properties

Objective. To examine the distributions of locality, flow, and coverage in Groupon_QL.
Methodology. We computed locality, flow, and coverage for each of the queries in Groupon_QL to obtain the distribution.
Results. The results are shown in Fig. 1.
Discussion. Each of the three distributions displays different characteristics. Locality is roughly evenly distributed in its range of $[0, 1]$. This is surprising, since we might expect the distribution to be significantly concentrated at higher values; however this is not so. This shows that queries are not, in general, significantly more local. Flow peaks around the value of 1. Since flow is an entropy, its unit is "number of bits." One bit of entropy describes a two state even distribution, and this is roughly the mode of the flow distribution. Coverage displays a falling mean-trend with increased values, except for a peak at coverage of one. The pearson correlation between the three quantities is very low, except for locality and flow which are negatively correlated [2].

6.3 Identifying Atypical Queries Accurately

Objective. To identify atypical queries in Groupon_QL, and investigate the accuracy of the proposed techniques.
Methodology. We isolated a set of 3000 previously untagged queries in Groupon_QL which had sufficient query volume (i.e. were queried more than 100 times in 6 months). We computed locality, flow, and coverage for each of these queries, and then used the framework of §4 to identify atypical queries. The thresholds to define the regions were: Region I (locality < 0.05 and flow > 3.5), Region II (locality < 0.05 and flow < 1.4) and Region III (locality > 0.7, flow < 1.4, and coverage < 0.05). This gave us 590 queries, out of which 120 were randomly sampled from each of the three regions to generate the test data set. Then, two experienced search analysts at Groupon were asked to tag each of these 120 queries as "maximize recall," "disambiguate" or "maximize precision," (resp.) depending on which region of atypicality the query lay.

[1] A frequently used technique is "search in:" where the user is presented with (for the example of apple) the choice to search in electronics or in food, in order to disambiguate.

[2] Correlation between coverage and locality is near zero, between locality and flow is -0.47 and between coverage and entropy is 0.11.

Table 1: Examples of atypical queries: causes, individual properties, and notes.

Cause of Atypicality	Query	Flow	Locality	Coverage	Clarifying Comments
low locality, high flow	romantic	5.27	0[a]	0.69	broad query
	organic	4.98	0[a]	0.50	broad query
	picnic	3.03	0.03	0.12	broad query
	amore	3.07	≈ 0	0.12	restaurant, dance class, spa, salon etc.
low locality, low flow	edison	0.97	≈ 0	0.33	hotel, also lighting goods
	solo	1.00	≈ 0	0.06	headphones, also italian restaurant
	allegria	0.90	≈ 0	0.04	popular restaurant, also hotel
	d-link	1.00	≈ 0	1.00	broadband routers, also surveillance cameras
high locality, low flow, low coverage[b]	squid lips	0.48	0.59	0.08	seafood restaurant
	meal replacement	1.00	0.90	0.28	specific intent
	olaf	1.31	0.55	0.17	stuffed toy for kids
	sound machine	0[c]	1[c]	0.04	sounds for babies
TYPICAL QUERY	pizza	1.08	0.88	0.82	medium flow, high locality and coverage

[a] Each category in supp(Q) has normalized count less than 0.1 [b] Eight of the top 20 low coverage queries correspond to brands

[c] $|\text{supp}\,(Q)| = 1$, hence flow(Q) $= 0$ and locality(Q) $= 1$ by definition

Results. Shown in Table 2

Discussion. Region III queries were most easily isolated by the two analysts, due to predominance of easily distinguishable brands, both local and global. The lowest agreement was found on Region II queries, and almost all the differing queries were assigned the label "maximize precision" by the analysts. This is both due to a certain amount of randomness in click activity and the presence of dual intent behind some seemingly specific queries. For instance, when people search for "Serta", the underlying intent is commonly assumed to be a mattress purchase, but it could also refer to a furniture purchase intent. Some Region I queries were assigned the label "maximize precision" due to specific keywords like "Sams Club" but the underlying click activity has very high flow because the intent cannot be localized to a definite category. Despite the shift in labels, the *common consensus amongst the analysts was that all the queries were atypical, and needed to be flagged for special care.*

Table 2: Analysts verdict (proportion of agreement) on top 120 queries tagged as atypical with our approach.

	Region I	Region II	Region III
Total Queries	40	30	50
Analyst A	0.75	0.83	1.00
Analyst B	0.68	0.57	0.86

6.4 Examples of Atypical Queries

Objective. To provide examples of atypical queries in Groupon_QL.

Methodology. From the atypical queries identified in §6.3, we show four of each of the three types.

Results. Twelve examples results are shown in Table. 1. Specifically, queries of type described in §4.1, §4.2, §4.3 are color coded orange, green, and blue, respectively.

Discussion. In table.

7. CONCLUSION

In this paper, we have introduced work on the important problem of structural understanding of queries in eCommerce in terms of specificity, breadth, and ambiguity. We identified three inherent properties of query-click distributions: geometric, statistical, and topological. We studied these properties, and showed that queries that have certain combinations of these properties are atypical in certain senses. By way of example, we showed three regions of atypicality. The framework that we established is flexible, and can be adapted by search analysts for specific applications.

Our work is being used in production to improve relevance of search results at Groupon. We would like to bring research questions from eCommerce retrieval to the attention of the broader retrieval community, and this work is a starting step in that direction.

8. DEDICATION

This work is dedicated to Shiva Kashid (d. July 1660) for his sacrifice in enabling Shivaji to escape from the siege of Panhala.

References

[1] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *Proc. SIGIR '06*, pages 390–397, New York, NY, USA, 2006. ACM.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. SIGIR '02*, pages 299–306, New York, NY, USA, 2002. ACM.

[4] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proc. WWW '07*, pages 581–590, New York, NY, USA, 2007. ACM.

[5] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Proc. CIKM '08*, pages 1419–1420, New York, NY, USA, 2008. ACM.

[6] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proc. CIKM '08*, pages 439–448, New York, NY, USA, 2008. ACM.

[7] J. Munkres. *Topology (2nd Edition)*. Pearson, 2 edition, 2000.

[8] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search. In *Proc. WWW '07*, pages 1169–1170, New York, NY, USA, 2007. ACM.

[9] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: Modeling queries with variation in user intent. In *Proc. SIGIR '08*, pages 163–170, New York, NY, USA, 2008. ACM.

[10] V. Vinay, I. J. Cox, N. Milic-Frayling, and K. Wood. On ranking the effectiveness of searches. In *Proc. SIGIR '06*, pages 398–404, New York, NY, USA, 2006. ACM.

[11] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proc. SIGIR '07*, pages 543–550, New York, NY, USA, 2007. ACM.